Layer-Level Progressive Transformer With Modality Difference Awareness for Multi-Modal Neural Machine Translation

Junjun Guo, Junjie Ye, Yan Xiang, and Zhengtao Yu

***Abstract*—Multi-modal neural machine translation (MNMT) aims to translate sentences from the source language into the target language with the aid of corresponding images. Unfortunately, there is a considerable modality gap between the semantic-related images and texts in terms of data form and semantic expression. How to fully incorporate visual information into texts to enhance the performance of machine translation is one of the critical issues for MNMT. However, the initial visual and textual features are generally extracted with their modality-specific models; Consequently, there is a considerable representation gap between images and texts. Most previous MNMT works prefer only to adopt the feature-level fusion strategies to learn multi-modal representation, while the modality representation gap is often ignored. To this end, this article proposes a progressive multi-modal Transformer (ProMul-Trans) with Modality Difference-Aware (MDA) to address the visual-to-textual fusion problem raised in MNMT. We first employ MDA to capture the modality-consistency information by taking visual and textual representations as inputs in each Transformer layer. Then a layer-level progressive fusion (Layer-ProFusion) strategy is adopted to progressively align visual and textual representations layer-by-layer to enhance machine translation performance. Experiment results on the Multi30 k dataset are conducted, and the results show that the proposed approach outperforms the compared state-of-the-art (SOTA) methods on English → German (En → De), English → French (En → Fr) and English → Czech (En → Cs) tasks. We release the code at https://github.com/JunjieYe-MMT/HierProMul-Trans.**

***Index Terms*—Layer-level progressive fusion, modality difference-aware, multi-modal neural machine translation, multi-modal transformer.**

Manuscript received 31 March 2022; revised 7 October 2022, 27 January 2023, 28 March 2023, and 15 May 2023; accepted 18 July 2023. Date of publication 2 August 2023; date of current version 11 August 2023. This work was supported in part by the National Key Research and Development Program of China under Grant 2020AAA0107904, in part by the National Natural Science Foundation of China under Grants 62241604 and 61866020, in part by the Natural Science Foundation Project of Yunnan Science and Technology Department under Grant 202301AT070444, in part by Yunnan Provincial Major Science and Technology Special Plan Projects under Grant 202103AA080015, and in part by Yunnan Key Research Projects under Grant 202203AA080004. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Preslav Nakov. *(Corresponding author: Yan Xiang.)*

The authors are with the Faculty of Information Engineering and Automation, Kunming University of Science and Technology, Kunming 650500, China, and also with the Yunnan Key Laboratory of Artificial Intelligence, Kunming University of Science and Technology, Kunming 650500, China (e-mail: guojjgb@163.com; yejunjie_cdx@163.com; sharonxiang@126.com; ztyu@hotmail.com).

Digital Object Identifier 10.1109/TASLP.2023.3301210

I. Introduction

MULTI-MODAL neural machine translation (MNMT) aims to translate source language sentences into target language sentences by incorporating visual information as additional context, which has received widespread attention over the past few years. A reasonable assumption of MNMT is that visual information helps improve machine translation [1], [2], [3], and many studies [4], [5], [6] have been carried out to demonstrate the benefits of using images for machine translation.

Unfortunately, there is a considerable modality gap between semantic-related images and texts in both data form and semantic expression. As shown in Fig. 1, objects, such as buses, bicycles, motorcycles, and people contained in the image are composed of pixels, whereas textual information is typically represented based on entities and their contexts, such as the words *"bus," "bicycle,"* and *"motorcyclist."* Thus, incorporating visual information into texts to enhance machine translation performance is a critical issues for MNMT [7]. There have been many attempts to explore the visual-to-textual fusion problem by adopting several feature fusion strategies, such as feature concatenation [6], [8], [9], [10], multi-modal gating [5], [6], [11], cross-attention [12], [13], [14], [15], [16], adaptive feature selection [17], [18], [19], [20], etc. In addition, other fusion strategies such as adversarial learning [21], [22] and reinforcement learning [23] also have been adopted for multi-modal fusion. However, most previous MNMT works only adopt feature-level fusion strategies to learn multi-modal representation, while ignoring the modality representation gap. The initial visual and textual features are generally extracted using pre-trained visual models and textual word embedding modules, respectively, resulting in a significant representation gap between images and texts.

Narrowing the representation gap and performing feature-level fusion are two coupled issues that need to be considered for multi-modal fusion. Typically, the closer the representation gap between images and texts, the more effectively visual and textual features can be fused in multi-modal feature space. Simultaneous awareness of representation gap and feature fusion might be one of the most effective ways to promote the network to learn better multi-modal representation.

To address the visual-to-textual fusion problem that arises in MNMT, this article presents a progressive visual-to-textual fusion framework based on MDA in the Transformer encoder. The proposed approach performs multi-modal fusion layer-by-layer,

En:A bus passes by a bicycle rider and a motorcyclist.
NMT:Ein bus fährt an einem motorradfahrer vorbei.
(*A bus drives past a motorcyclist.*)

Fig. 1. En → De multi-modal neural machine translation example.

progressively incorporating visual information into texts to enhance machine translation performance. Compared to existing works, the major contributions of our article are three-fold.

- We propose an MDA-enhanced ProMul-Trans approach to address the multi-modal fusion problem that arises in MNMT, with the goal of improving machine translation performance by effectively incorporating visual information into the translation process.
- Our approach adopts a layer-level progressive multi-modal fusion strategy, called Layer-ProFusion, which is based on MDA and aimed at encouraging the encoder to learn better multi-modal representation. Specifically, the MDA module captures overlapping semantic information between texts and images at each encoder layer. We then employ the MDA-based Layer-ProFusion strategy to progressively incorporate visual features into texts, thereby narrowing the gap between the two modality in the multi-modal representation space.
- Extensive experimental results on the Multi30 K benchmark dataset demonstrate that our proposed model outperforms other SOTA MNMT approaches and significantly improves machine translation performance on the English-German, English-French, and English-Czech translation tasks.

II. Related Work

MNMT has drawn much attention over the past few years, which has become one of the most active fields in natural language processing (NLP). Learning better representations in multi-modal common feature space is one of the significant challenges for MNMT. In the following, we provide an overview of visual-to-textual fusion approaches for MNMT.

A. RNN-Based MNMT Model

Earlier multi-modal fusion methods are mainly presented in RNN-based seq2seq frameworks. Global visual features extracted from the pre-trained CNN models were used to initialize the hidden states of the RNN encoder and/or decoder [18], [24], [25], or to augment the textual features as an additional inputs [18]. Although these methods could improve the performance of machine translation, visual features are not aligned with textual features. To better align visual and textual semantic features, Caglayan et al. [26], [27] leveraged a multi-modal attention mechanism to simultaneously pay attention to images and their corresponding texts; Calixto et al. [19] adopted two modality-specific attention mechanisms to align both visual and textual features for MNMT. Delbrouck and Dupont [28] proposed a local visual attention mechanism to align local visual features with textual features to enhance the performance of machine translation.

B. Transformer-Based MNMT Model

With the development of machine translation techniques, most recent MNMT approaches are presented based on Transformer structure [29]. We summarize the existing multi-modal fusion strategies in the following three aspects: 1) Cross-modal interactive attention mechanism, Zhao et al. [16] utilized object detection features with an additional region-dependent attention mechanism to fuse regional visual and textual features; Nishihara et al. [30] presented a supervised cross-modal attention module to align textual and visual features; Song et al. [14] employed a co-attention graph updating module at each Transformer encoder layer to align multi-modal features. 2) Feature concatenating methods, Yao et al. [8] used a multi-modal Transformer to align both visual and textual features; Takushima et al. [9] concatenated both global visual features and textual features as multi-modal features for MNMT; Li et al. [6] directly concatenated textual and visual representations as the multi-modal representation to preserve fine-grained features and avoid confounding modality-specific features. 3) The gating fusion methods, Yin et al. [5] proposed a graph-based MNMT approach to extract multi-model features through a text-image gating attention mechanism; Lin et al. [11] adopted a gating mechanism to fuse visual features extracted by a dynamic context-guided capsule network; Li et al. [6] used a gating fusion approach to resolve ambiguous word translation problems; Zhao et al. [15] proposed a word-region alignment-guided approach to align textual and visual features for MNMT. However, most MNMT works only focus on feature-level fusion based on the initial visual and textual features extracted from modality-specific networks in their corresponding feature space. The modality representation gap between images and texts is often ignored in the above MNMT approaches.

III. Methodology

MNMT faces more significant challenges due to the modality gap between images and texts than other text-only NMT approaches. While many current MNMT approaches are practical, they may not effectively bridge the modality gap between images and texts, thereby limiting the further improvement of machine translation performance. In this article, we propose an MDA-enhanced ProMul-Trans encoder that adopts a layer-level progressive multi-modal fusion strategy to enhance machine translation performance. The framework of the ProMuli-Trans encoder is illustrated in Fig. 2.

There are several stacked encoder layer in the ProMul-Trans encoder, each encoder layer comprises four sub-layers: 1) Feature extraction layer; 2) The MDA module; 3) Cross-modal interaction module; and 4) The layer-ProFusion module.

A. Feature Extraction

Without loss of generality, denote by $x_k = \{x_1^k, \ldots, x_n^k\}$ and z_k as the source sentence and the image of the k-th data-pair,

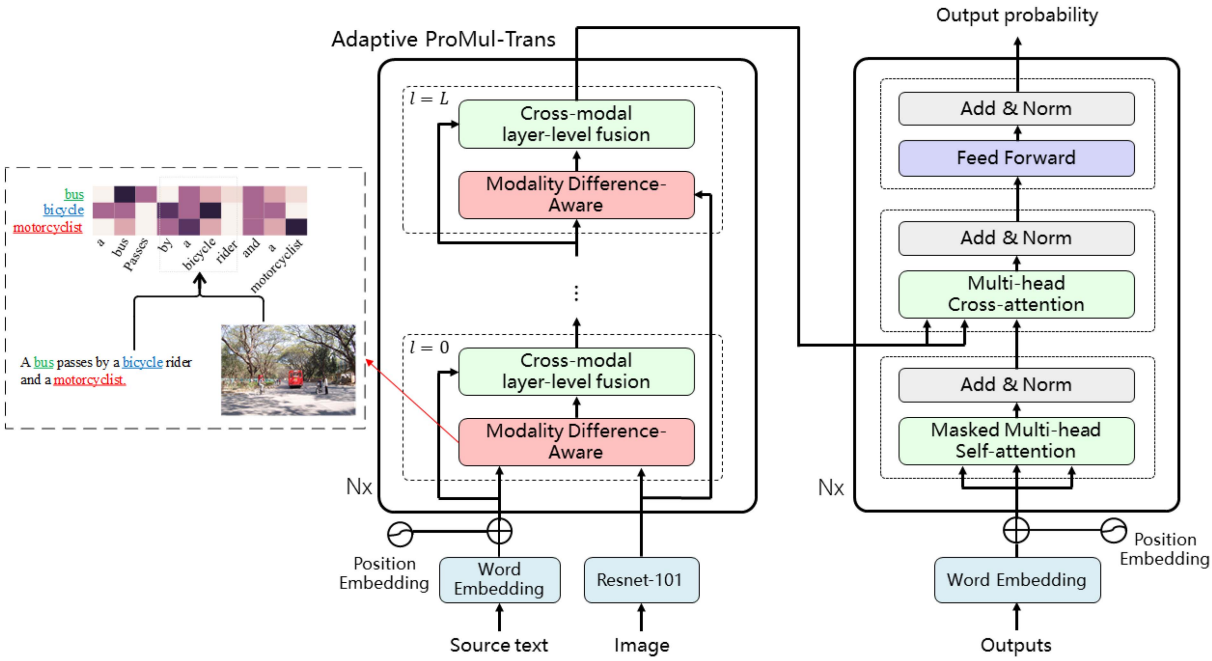

Fig. 2. Architecture of our proposed MNMT.

respectively, where n is the length of x_k. The sentence words are embedded via traditional embedding layer with position embedding, and image features are extracted by the pre-trained Resnet-101 model [31]. Formally, the textual representation E_k^x and visual representation E_k^z are calculated as follows:

$$\mathrm{E}_k^x = \mathrm{emb}_x(x_k) \tag{1}$$

$$\mathrm{E}_k^z = \mathrm{emb}_z(z_k) \tag{2}$$

where, emb_x is the textual embedding layer with both word embedding and position embedding, emb_z is the visual feature extraction layer with Resnet-101, $\mathrm{E}_k^x \in R^{n\times d_1}$ and $\mathrm{E}_k^z \in R^{7\times 7\times d_2}$.

B. The MDA Module

The modality gap between images and texts is one of the critical factors affecting multi-modal fusion. Awareness of modality differences might be the most direct way to quantify the visual-to-textual gap in multi-modal common semantic space. It would be helpful to enhance the performance of machine translation. To this end, a novel MDA module is proposed to dynamically quantify the modality gap between images and texts in each encoder layer. The MDA module can promote the network to learn the modality-consistency representation based on the gated fusion mechanism by taking both visual and textual features as inputs. The proposed MDA structure is visualized in Fig. 3. Formally, the extracted modality-consistency representation $\mathrm{C}_k^{\mathrm{con},l-1}$ are computed as follows,

$$\sigma_k^o = \mathrm{Sigmoid}(\mathbf{W}_o \cdot (\mathrm{C}_k^{x,l-1} \,||\, (\mathbf{W}_z \mathrm{C}_k^z))) \tag{3}$$

$$\mathrm{C}_k^{\mathrm{con},l-1} = \sigma_k^o \odot \mathbf{W}_x \mathrm{C}_k^{x,l-1} \tag{4}$$

where, C_k^z and $\mathrm{C}_k^{x,l-1}$ denote the visual representation and the $l-1$-th layer textual representation, respectively, l is the Transformer layer index and $\mathrm{C}_k^{x,0} = \mathrm{E}_k^x$ when l=0, $\mathrm{C}_k^z = \mathrm{E}_k^z$; $||$ denotes the feature concatenation operation in the last dimension, the multi-modal gate $\sigma_k^o \in \mathrm{R}^{n\times d_1}$, $\odot$ is the point-wise multiplication operation, $\mathbf{W}_x$, $\mathbf{W}_z$, $\mathbf{W}_o$ are trainable parameters, $\mathrm{C}_k^{\mathrm{con},l-1} \in \mathrm{R}^{n\times d_1}$ denotes the learned modality-consistency representation between images and texts.

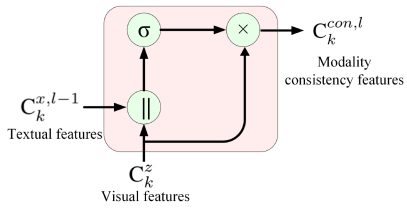

Fig. 3. Diagrams of the MDA module.

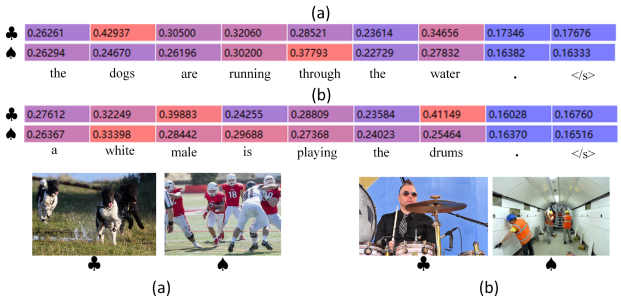

Fig. 4. Visualization of σ_k^0 in MDA. The club symbol represents the image paired with the text, the spade symbol represents a randomly selected image.

Study of MDA: Inspired by [32], we conduct an experiment to evaluate MDA's ability to capture modality-consistency information. In Fig. 4, we have visualized the sigmoid function of two samples. The club means that the image is semantically

related to the sentence, while the spade means that the image does not match the sentence (we randomly selected an image loosely related to the sentence from the dataset). The results show that 1) MDA can capture visual-related words (such as "artist," "dogs," and "male") for the matched text-image pair; 2) when the image is loosely related to the sentence, MDA assigns higher weights to unimportant words (such as "through," and "white"); and 3) MDA can still capture visual-related words even when the image is loosely related to the sentence. These results highlight the effectiveness of MDA's gating mechanism in capturing overlapping semantic information between texts and images.

Then the well-known mixup strategy is employed to mix the modality-consistency features $\mathrm{C}_k^{\mathrm{con},l-1}$ with the textual features $\mathrm{C}_k^{x,l-1}$ to learn the visual-enhanced multi-modal representation $\mathrm{C}_k^{m,l}$, which has been widely explored in CV [33], [34], [35] and NLP tasks [36], [37], [38], [39],

$$\mathrm{C}_k^{m,l} = \varphi \cdot \mathrm{C}_k^{\mathrm{con},l-1} + (1-\varphi) \cdot \mathrm{C}_k^{x,l-1} \tag{5}$$

where φ is a scalar mixing hyper-parameter leveraged to balance both the modality-consistency feature and the textual feature, $\mathrm{C}_k^{m,l}$ is the mixed multi-modal representation of the k-th data-pair.

C. Cross-Modal Interaction Module

Then the cross-modal interaction module is employed in each Transformer layer based on MDA to promote multi-modal alignment. We leverage the visual-enhanced multi-modal representation $\mathrm{C}_k^{m,l}$ as the query matrix and the textual representation $\mathrm{C}_k^{x,l-1}$ as the key/value matrix. The cross-modal interaction module can be represented as follows:

$$\mathrm{H}_k^l = \mathrm{Multihead}\left(\mathrm{C}_k^{m,l}, \mathrm{C}_k^{x,l-1}, \mathrm{C}_k^{x,l-1}\right) \tag{6}$$

$$= \mathrm{Concat}\left(\mathrm{head}_k^1, \ldots, \mathrm{head}_k^M\right) \tag{7}$$

where, Multihead $(\cdot)$ is a multi-head attention layer, M denotes the number of heads. Formally, the head attention is computed as follows:

$$\mathrm{head}_k^{c\in[1,M]} = \sum_{j=1}^{n} \alpha_{ij} \left(\mathbf{W}_{k,c}^V \mathrm{C}_k^{x_j,l-1}\right) \tag{8}$$

where, the weight coefficient α_{ij} is calculated by the softmax function:

$$\alpha_{ij} = \mathrm{softmax}\left(\frac{(\mathbf{W}_{k,c}^Q \mathrm{C}_k^{m_i,l})(\mathbf{W}_{k,c}^K \mathrm{C}_k^{x_j,l-1})^{\mathrm{T}}}{\sqrt{d}}\right) \tag{9}$$

where α_{ij} is the dot-product attention matrix, and i, j represent the index of visual-enhanced multi-modal tokens and textual tokens, respectively. $\mathbf{W}_{k,c}^V$, $\mathbf{W}_{k,c}^Q$, $\mathbf{W}_{k,c}^K$ are trainable parameter matrices. We adopt the attention mechanism to represent the similarity of textual features and the mixed multi-modal features.

Similar to conventional Transformer-model architecture [40], a position-wise Feed-Forward network (FFN) is then employed as follows:

$$\mathrm{C}_k^{x,l} = \mathrm{FFN}\left(\mathrm{H}_k^l\right) \tag{10}$$

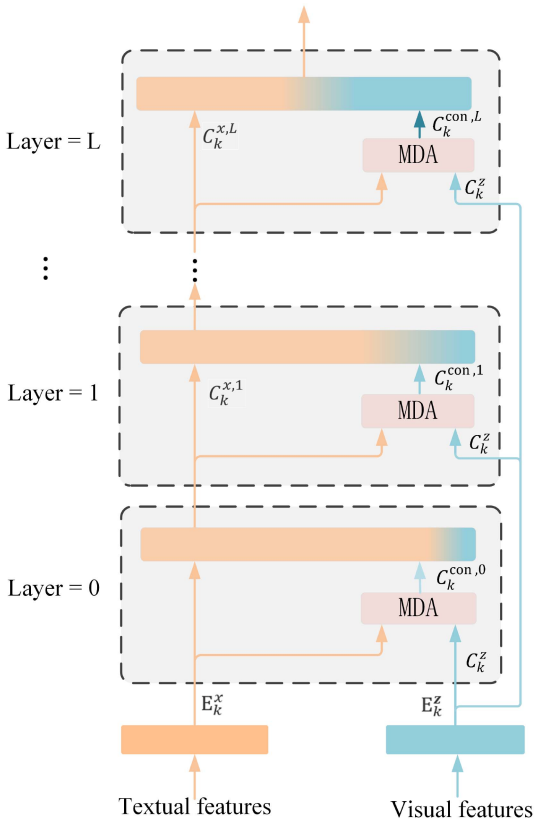

Fig. 5. Layer-level progressive fusion strategy in ProMul-Trans encoder.

The output of the l-th transformer encoder layer $C_k^{x,l}$ is then sent to the next transformer layer to further obtain the interactive multi-modal representation. Denote by L the index of the last transformer encoder layer, and the output of the transformer encoder $C_k^{x,L}$ is fed into the transformer decoder for target sentence generation. Moreover, our model also adds the residual connection between each layer and layer normalization in each encoder layer.

D. Layer-ProFusion Strategy

Due to the modality gap between images and texts, it is generally challenging to promote multi-modal alignment by directly mapping visual features into textual semantic space. The unaligned visual and textual features will lead to semantic information inconsistency in the learned multi-modal representation, affecting machine translation. To better narrow the multi-modal gap, a Layer-ProFusion strategy is adopted in the ProMul-Trans encoder, which can gradually incorporate visual information into texts layer-by-layer with the guidance of modality-consistency information. More concretely, when layer index l increases, image information is more incorporated into texts. The proposed Layer-ProFusion strategy is illustrated in Fig. 5.

$$\mathrm{C}_k^{m,l} = \beta \cdot \mathrm{C}_k^{\mathrm{con},l-1} + (1-\beta) \cdot \mathrm{C}_k^{x,l-1} \tag{11}$$

$$\beta = \gamma \cdot \sqrt{(l+1)} \tag{12}$$

where γ is the hyper-parameter, l is the Transformer layer index, β is leveraged to balance both modality-consistency feature $C_k^{\text{con},l-1}$ and textual feature $C_k^{x,l-1}$, $C_k^{m,l}$ is the fused multi-modal feature in the l-th transformer layer.

Thus, modality-consistency feature extraction and multi-modal fusion are interactively employed layer-by-layer to progressively incorporate visual information into texts in the ProMul-Trans encoder. Then, the classic Transformer decoder generates the target word sequentially by taking the output of our proposed multi-modal Transformer encoder and the target sentence as inputs.

IV. Experiment

We evaluate the performance of our model with the other SOTA MNMT approaches on En → De, En → Fr and En → Cs translation tasks, and we conduct experiments on four test sets: 1) the Test2016, 2) the Test2017, 3) the MSCOCO and 4) the Test2018. We first describe the implementation details and the experimental setup. Then we compare our approach with baselines with detailed analysis.

A. Datasets

Text: We carry out all experimental results on the Multi30K[1] dataset [41]. The training, validation and testing sets contain 29,000, 1,014 and 1,000 text-image pairs, respectively. All sentences in train, valid and test data sets have been pre-processed to lowercase, normalized punctuation and sentence tokenization. We directly use the pre-processed sentence pairs [5] via byte pair encoding [42] segmentation with 10,000 merge operations. The vocabulary sizes of each language pair are 8,503→9,388 tokens for the En → De translation task, 8,503→8,745 tokens for the En → Fr translation task and 8,503→9,426 tokens for the En → Cs translation task. Each image is paired with an image description expressed by the original English sentence and three target language sentences (German, French and Czech). The experiments are conducted on four types of test sets: 1) The Test2016 test set contains 1,000 sentence pairs; 2) The Test2017 test set contains 1,000 instances in WMT2017 with more difficult source sentences to understand and translate; 3) The MSCOCO test set contains 461 sentences with ambiguous verbs and encourages using images for disambiguation; and 4) The Test2018 test set contains 1,071 instances with ambiguous verbs.

Image: Visual features are extracted through the pre-trained Resnet-101 [31], the spatial features are 7x7x2048-dimensional vectors with 49 local spatial region features of the image.

B. Settings

We implement our proposed model based on the Transformer framework [40]. We try different Transformer model sizes (Big, Base, Small, and Tiny Transformers, see Table I) to find the suitable parameter settings for the Multi30 K dataset. It is worth noting that most of the existing MNMT approaches are based on Base [15], [44] or Small Transformer [45]. Compared with most previous works using a 6-layer encoder-decoder, we only stack 4 layers in both encoder and decoder. Concretely, the max tokens are set to 4,096, the warmup-update is set to 2,000, and the label smoothing value is 0.1. We use Adam optimizer with $\beta_1, \beta_2 = (0.9, 0.98)$. We adopt four heads here, and the dropout is set to 0.3. The learning rate is set to 0.005 and 0.008 for the En → De and En → Fr translation tasks, respectively. We stop the training when the BLEU score has not improved within 15 epochs on the validation set. We train our models on a single GTX 3090 GPU with fp16. Similar to most previous works, the metrics 4-gram BLEU [46] and METEOR[2] [47] are employed to evaluate the performance of our model and report the average scores over three runs.

TABLE I
Model Configurations for Big, Base, Small, and Tiny Transformer [43]

Transformer Sizes				
Model component	Big	Base	Small	Tiny
encoder/decoder FFN dimension	4096	2048	1024	256
encoder/decoder attention heads	16	8	4	4
encoder/decoder embedding dimension	1024	512	512	128
Number of encoder/decoder layers	6	6	6	4

C. Baseline Models

To verify the advantages of our proposed model, we compare the performance of the following recent SOTA MNMT models on the En → De, En → Fr and En → Cs translation tasks,

- VMMT [48]: A GRU-based MNMT model that incorporates image context learnted by a latent variable model.
- VAG-NMT [17]: A visual-textual attention mechanism is employed to map textual and visual features into the multi-modal shared space.
- Del+Obj [44]: The image and target sentence are jointly trained in the decoder to generate good first-draft translations.
- DCCN [11]: A novel dynamic context-guided capsule network is proposed to guide visual feature extraction to improve machine translation performance.
- MNMT+SVA [30]: A supervised visual attention mechanism is proposed to capture the text-related visual regions for multi-modal machine translation.
- OVC+L_v [13]: An object-level visual context modeling framework is built to efficiently explore and capture visual information to guide machine translation.
- WRA-guided [15]: A word-region alignment-guided approach is proposed to establish semantic correlations between textual and visual features.

To ensure a fair comparison and demonstrate the superiority of our proposed model, we reproduce several recent SOTA MNMT approaches using the nearly identical parameter settings and the same framework (the fairseq tool).

[1][Online]. Available: https://github.com/multi30 k/dataset

[2][Online]. Available: http://www.cs.cmu.edu/alavie/METEOR/

TABLE II
COMPARISON RESULTS ON THE EN → DE TRANSLATION TASK ON THE MULTI30 K DATASET

Model	Multi30K En→De						
	Parameter	Test2016		Test2017		MSCOCO	
		BLEU	METEOR	BLEU	METEOR	BLEU	METEOR
Existing MNMT Systems							
VMMT [48]	-	37.7	56.0	30.1	49.9	25.5	44.8
VAG-NMT [17]	-	-	-	31.6	52.2	28.3	48.0
Del+Obj [44]	-	38.0	55.6	-	-	-	-
DCCN [11]	17.1M	39.7	56.8	31.0	49.9	26.7	45.7
MNMT+SVA [30]	-	39.9	58.1	-	-	-	-
OVC+L_v [13]	11.3M	-	-	32.4	52.3	28.6	48.0
WRA-guided [15]	-	39.3	58.3	32.3	**52.8**	28.5	48.5
Comparison Results of Different Transformer Frameworks							
Transformer-Big	196.8M	36.07	54.36	27.22	47.20	24.66	44.03
Transformer-Base	54.4M	38.60	57.24	31.66	51.00	28.81	47.95
Transformer-Small	41.7M	39.88	57.93	32.08	51.11	28.68	48.27
Transformer-Tiny	3.8M	40.96	58.35	32.59	51.21	29.16	48.37
Our MNMT Systems Based on Transformer-Tiny							
Doubly-ATT [49] †	4.1M	41.44	59.08	33.15	52.34	29.22	48.41
Multimodal self-att [8] †	3.8M	41.50	58.52	32.51	51.33	29.10	48.48
Gated Fusion MNMT [5] †	3.8M	41.58	58.88	33.01	51.90	30.04	48.95
Our model	3.8M	**42.00**	**59.43**	**34.08**	52.54	**30.38**	**49.60**

† means to reproduce previous MNMT methods based on our framework. Best results are highlighted in bold.

- Gated Fusion MNMT [5]: An efficient multi-modal To ensure a fair comparison and demonstrate the superiority of our proposed model, we reproduce several recent SOTA MNMT approaches using the nearly identical parameter settings and the same framework (the fairseq tool).
- Multimodal self-att [8]: A multi-modal transformer is designed to extract the most relevant visual information and improve machine translation performance by concatenating textual and visual features as the inputs of the multi-modal cross-attention module.
- Doubly-ATT [49]: An additional doubly-attention sublayer is inserted between the source-target attention and the feed-forward sublayers in the decoder.

D. Results on the En → De Translation Task

Table II reports the main translation results of our model and other SOTA models on the En → De translation task. Our proposed model outperforms most existing baseline models and only requires a small number of training parameters. Compared with the other models in Table II, we draw the following interesting conclusions:

1) Compared With Existing MNMT Approaches: Our model outperforms existing SOTA models, and improves BLEU and METEOR scores by 2∼3 points on most test sets. It confirms that our proposed method could effectively use visual features to improve machine translation.

2) Compared With Text-Only NMT: Our MNMT model significantly outperforms text-only NMT, achieving a 1∼2 point score improvement on all test sets. The experiment results show the effectiveness of visual information in improving the quality of machine translation.

3) Compared With Other Transformer Tiny-Based Methods: Our proposed method outperforms the other SOTA methods, which confirms the effectiveness of the proposed Lay-ProFusion strategy for MNMT.

E. Results on the En → Fr Translation Task

To verify the generalization of our proposed model, we also conduct experiments on the En → Fr translation task, as shown in Table III. We also implement Transformer with various sizes and reproduce recent SOTA methods on the En → Fr task. We can observe the following conclusions:

First, our model outperforms existing MNMT models on most of the test sets, which is consistent with the result of the En → De task. Furthermore, compared with text-only models, the Transformer Tiny-based model achieves higher scores than the other Transformer frameworks.

Second, compared with current competitive methods in the same Transformer-Tiny framework, our method still achieves the best value on most test sets. Additionally, compared with text-only NMT, our MNMT model significantly improves machine translation performance, which is consistent with the En → De task. Experimental results suggest that our model can effectively utilize image information to improve machine translation performance.

F. Results on the En → Cs Translation Task

To further verify the effectiveness and generality of our MNMT model, we conduct experiments on the En → Cs [2]

TABLE III
COMPARISON RESULTS ON THE EN → FR TRANSLATION TASK ON THE MULTI30 K DATASET

Model	Multi30K En→Fr					
	Test2016		Test2017		MSCOCO	
	BLEU	METEOR	BLEU	METEOR	BLEU	METEOR
Existing MNMT Systems						
VAG-NMT [17]	-	-	53.8	70.3	45.0	64.7
Del+Obj [44]	59.8	74.4	-	-	-	-
DCCN [11]	61.2	76.4	54.3	70.3	45.4	65.0
OVC+L_v [13]	-	-	54.2	70.5	45.2	64.6
WRA-guided [15]	61.8	76.3	54.1	70.6	43.4	63.8
Comparison Results of Different Transformer Frameworks						
Transformer-Big	56.74	72.49	49.45	67.42	41.73	62.51
Transformer-Base	59.15	75.49	51.60	70.44	42.82	65.15
Transformer-Small	59.65	75.36	52.80	71.76	43.98	65.61
Transformer-Tiny	60.33	75.64	53.45	71.57	43.61	65.72
Our MNMT Systems Based on Transformer-Tiny						
Doubly-ATT [49] †	60.94	75.99	53.63	71.56	44.78	65.35
Multimodal self-att [8] †	61.44	75.77	**54.56**	71.62	44.59	65.08
Gated Fusion MNMT [5] †	61.24	76.26	54.15	71.77	44.29	64.91
Our model	**62.36**	**77.20**	54.09	**72.09**	**46.48**	**66.71**

† means to reproduce previous MNMT methods based on our framework for MNMT. Best results are highlighted in bold.

TABLE IV
COMPARISON RESULTS ON THE EN-CS TRANSLATION TASK

Model	En→Cs			
	Test2016		Test2018	
	BLEU	METEOR	BLEU	METEOR
Our model	**33.80**	**32.49**	**29.94**	**30.03**
Transformer-Tiny	32.70	32.34	27.62	29.03
Doubly-ATT [49] †	33.25	32.28	29.12	29.87
Multimodal self-att [8] †	33.12	32.01	28.75	29.51
Gated Fusion MNMT [5] †	33.77	32.24	29.43	29.41
DeepLearnXMU [20] ‡	31.6	29.7	-	-

† means to reproduce previous MNMT methods based on Transformer-Tiny. ‡ denotes the existing experimental results.

TABLE V
ABLATION STUDY OF MDA ON THE EN → DE, EN → FR AND EN → CS TRANSLATION TASKS

#	Our model	Test2016		Test2017		MSCOCO		Test2018	
		BLEU	METEOR	BLEU	METEOR	BLEU	METEOR	BLEU	METEOR
				Multi30K En→De					
1	Our model	**42.00**	**59.43**	**34.08**	**52.54**	**30.24**	**49.60**	-	-
2	Remove MDA	41.12	58.77	33.16	52.02	29.38	48.59	-	-
				Multi30K En→Fr					
3	Our model	**62.36**	**77.20**	**54.09**	**72.04**	**45.49**	**66.42**	-	-
4	Remove MDA	61.48	76.39	53.73	71.92	44.74	65.90	-	-
				Multi30K En→Cs					
5	Our model	**33.80**	**32.49**	-	-	-	-	**29.94**	**30.03**
6	Remove MDA	33.43	32.39	-	-	-	-	29.02	29.59

translation task. The En → Cs translation task is the third shared translation task of the International Machine Translation Competition, which is a very challenging task with many ambiguous words. As reported in Table IX, our proposed model outperforms previous methods and achieves the SOTA results on the En → Cs translation task. This further confirms the effectiveness and generality of our proposed model in different languages.

G. Ablation Study

To verify the effectiveness of our proposed method, we conducted ablation studies on three translation tasks based on Transformer-Tiny. 1) ablation study on MDA, 2) ablation study on Layer-ProFusion strategy, and 3) validity of image information.

1) Ablation Study on MDA: We first conduct an ablation study on MDA to verify its contribution, as shown in Table V. Experiment results show that removing MDA leads to a significant performance decline on three test sets. The results confirm the effectiveness of MDA to improve machine translation.

2) Ablation Study on Layer-ProFusion: The impact of layer-level progressive multi-modal fusion strategy is discussed in this sub-section: The effect of different fusion ratios, the effect of the multi-modal feature mixing strategy, and the effect of the layer-level fusion mechanism. We can obtain the following interesting conclusions:

Effect of different fusion ratios: We can observe that the experimental results of our proposed model achieve higher machine translation scores when $\gamma = 0.2$ on the En → De and En → Fr translation tasks. So we choose $\gamma = 0.2$ as our experiment setting.

Effect of the progressive fusion strategy: Comparing the Layer-ProFusion strategy with the fixed scale fusion on two translation tasks, the experiment results of Table VI show that the progressive fusion yields additional gains on all test sets with an average gain of around 0.5 points. Furthermore, the progressive

TABLE VI
ABLATION STUDY ON PROGRESSIVE FUSION WITH DIFFERENT γ, φ VALUES ON THE EN $\rightarrow$ DE, EN $\rightarrow$ FR AND EN $\rightarrow$ CS TRANSLATION TASKS

		Test2016		Test2017		MSCOCO		Test2018	
	Fusion strategy	BLEU	METEOR	BLEU	METEOR	BLEU	METEOR	BLEU	METEOR
Multi30K En→De									
$\gamma, \varphi = 0.1$	Progressive	41.31	59.03	33.35	52.20	30.07	49.47	-	-
	Fixed Scale	41.25	58.69	32.73	51.84	29.97	48.96	-	-
$\gamma, \varphi = 0.2$	Progressive	**42.00**	**59.43**	**34.08**	**52.54**	30.24	**49.60**	-	-
	Fixed Scale	40.84	59.23	32.48	52.2	29.65	49.24	-	-
$\gamma, \varphi = 0.3$	Progressive	40.96	58.70	33.49	52.05	29.78	49.09	-	-
	Fixed Scale	40.59	58.70	32.93	51.74	29.96	48.73	-	-
$\gamma, \varphi = 0.4$	Progressive	41.49	59.11	32.78	51.86	30.09	49.58	-	-
	Fixed Scale	40.99	58.66	32.74	51.83	29.77	49.43	-	-
$\gamma, \varphi = 0.5$	Progressive	41.39	58.95	32.83	51.94	**30.38**	49.37	-	-
	Fixed Scale	41.06	58.74	33.11	52.09	29.85	48.89	-	-
Multi30K En→Fr									
$\gamma, \varphi = 0.1$	Progressive	61.67	76.83	54.07	72.01	45.46	66.28	-	-
	Fixed Scale	61.44	76.15	53.89	71.52	45.57	65.91	-	-
$\gamma, \varphi = 0.2$	Progressive	**62.36**	**77.20**	**54.09**	72.04	45.49	66.42	-	-
	Fixed Scale	61.52	76.34	53.95	71.85	45.79	66.49	-	-
$\gamma, \varphi = 0.3$	Progressive	61.32	76.14	53.88	72.06	44.55	65.36	-	-
	Fixed Scale	61.62	76.28	53.26	71.32	44.84	65.46	-	-
$\gamma, \varphi = 0.4$	Progressive	61.99	76.47	54.03	**72.09**	**46.48**	**66.71**	-	-
	Fixed Scale	61.68	76.19	53.92	71.78	45.53	66.38	-	-
$\gamma, \varphi = 0.5$	Progressive	61.05	75.72	53.62	71.18	44.43	65.32	-	-
	Fixed Scale	60.89	75.76	53.84	70.86	44.83	64.73	-	-
Multi30K En→Cs									
$\gamma, \varphi = 0.2$	Progressive	**33.80**	**32.49**	-	-	-	-	**29.94**	**30.03**
	Fixed Scale	33.71	32.47	-	-	-	-	29.11	29.78

TABLE VII
ABLATION STUDY ON LAYER-LEVEL FUSION ON THE EN $\rightarrow$ DE AND EN $\rightarrow$ FR TRANSLATION TASKS

Our Model	Test2016		Test2017		MSCOCO	
	BLEU	METEOR	BLEU	METEOR	BLEU	METEOR
Multi30K En→De						
Our model	**42.00**	**59.43**	**34.08**	**52.54**	**30.24**	**49.60**
Remove C_k^z in the 3rd layer	41.34	58.87	33.12	52.09	29.41	49.03
Remove C_k^z in the 2nd and 3rd layers	40.97	58.78	33.02	51.99	29.65	48.80
Remove C_k^z in the 1st, 2nd and 3rd layers	40.69	58.53	32.85	51.42	28.09	47.86
Multi30K En→Fr						
Our model	**62.36**	**77.20**	**54.09**	**72.04**	**45.49**	**66.42**
Remove C_k^z in the 3rd layer	61.85	76.69	53.90	71.63	45.16	66.01
Remove C_k^z in the 2nd and 3rd layers	61.38	76.27	53.94	71.66	44.84	65.96
Remove C_k^z in the 1st, 2nd and 3rd layers	61.33	75.93	54.21	71.59	44.85	65.95

fusion strategy outperforms the fixed scale fusion strategy even under different γ parameter settings on the En $\rightarrow$ De task, which is consistent with the results on the En $\rightarrow$ Fr and En $\rightarrow$ Cs translation tasks. These findings indicate that progressive fusion can significantly improve the performance of MNMT models.

Effect of the layer-level fusion mechanism: The impact of the layer-level progressive fusion mechanism is illustrated in Table VII. The machine translation scores decrease gradually when the visual features are gradually removed from the top down, consistent with the findings in the En $\rightarrow$ Fr task. Experimental results validate the effectiveness of the layer-level progressive fusion mechanism.

Furthermore, we visualize the cosine similarity between text representation $C_k^{x,l}$ and image representation C_k^z in each layer of the MDA-enhanced ProMul-Trans encoder (where l denotes the layer index, and we choose $l = 0, 1, 2, 3$ in the manuscript), as shown in Fig. 6. The horizontal axis indicates the index of the image-text samples, and the vertical axis represents the similarity between images and texts. The results show that our approach successfully aligns the text and image representations layer-by-layer, indirectly indicating the ability of our model to align the two modalities.

3) Validity of Image Information: Inspired by Elliott [21], we further examine the utility of images in machine translation by

TABLE VIII
VALIDITY OF IMAGE INFORMATION ON THE EN → DE AND EN → FR TRANSLATION TASKS

#	Our model	Test2016		Test2017		MSCOCO	
		BLEU	METEOR	BLEU	METEOR	BLEU	METEOR
	Multi30K En→De						
1	Our model	**42.00**	**59.43**	**34.08**	**52.54**	**30.24**	**49.60**
2	Replace images with blank images	41.29	58.98	33.07	52.11	29.66	49.07
3	Replace images with random images	40.66	58.44	32.75	51.75	29.01	48.05
	Multi30K En→Fr						
4	Our model	**62.36**	**77.20**	**54.09**	**72.04**	**45.49**	**66.42**
5	Replace images with blank images	61.03	76.35	54.03	71.98	44.91	65.86
6	Replace images with random images	60.36	75.28	52.24	70.55	44.06	64.78

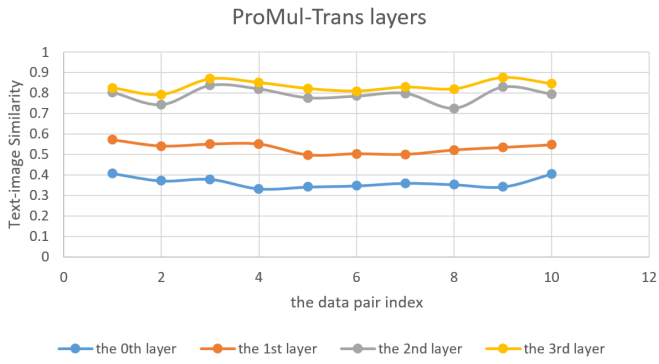

Fig. 6. Visualization of the similarity of images and texts in different ProMul-Trans layers.

adversarial evaluation. As shown in rows 2 and 5 of Table VIII, when we replace the input images with blank images, the performance of our model drops significantly. The results confirm the hypothesis of our work that visual information is crucial for enhancing machine translation performance. As shown in rows 3 and 6 of Table VIII, when we replace input images with randomly selected images from the dataset, the performance is worse than that of the input blank images. The underlying reason may be that the randomly selected visual features are noisy, interfering with the translation.

H. Case Study

Fig. 7 depicts the translation of two cases in Test2017 test set. Colors highlight improvement. In these examples, all baselines do not correctly translate the low-frequency nouns or noun phrases *"spaghetti"* and *"log cabin."* However, our model can generate them accurately (*"spaghetti"* appear one time, and *"holzhütte"* appears two times in the Multi30 K training set). The case study indicates that our proposed MNMT model can better use visual information to improve translation performance.

Inspired by [29], we then generated a "heat map" to illustrate the correlation between text and image. This was accomplished by overlapping the heat map onto the image for a given image-text pair, as shown in Fig. 8. The heat map results validate that the MDA-enhanced approach successfully mapped significant tokens onto their respective image regions, such as "lady," "white," and "tennis racket." These heat map visualizations provide additional evidence of the proposed approach's effectiveness in aligning the two modalities.

I. Human Analysis

To quantitatively demonstrate the effectiveness of the proposed approach, we conduct two types of fine-grained human analysis on the En-De and En-Fr translation tasks. We measure the quality of our translations by calculating the average fidelity score on several randomly selected sets of 200 samples. To ensure objective evaluation, we invite two Ph.D. students to independently assess the translation results using fidelity, which involves comparing the machine-translated output to the reference translation. We define the average fidelity score as follow:

$$\mathrm{F}_s = 100 * \frac{1}{N} \sum_i \frac{S_i}{S} \tag{13}$$

where, i denotes the student index, and N is the total number of Ph.D. students; S_i and S represent the fidelity score given by student i and the total fidelity score for the translated text.

Table IX shows the average fidelity scores for two translation tasks on six randomly selected test sets of 200 samples. The results show that 1) our model significantly improves fidelity scores compared to text-only NMT approach; And 2) there is a significant decrease in translation performance when the original images are replaced with randomly selected images (following the similar experiment setting in Fig. 4). Human analysis of the En-De and En-Fr translation tasks confirms the effectiveness of the proposed approach for machine translation.

As we know, images contain significant information in the form of entities and poses. To conduct a more thorough analysis of the translation quality, we perform a quantitative human analysis of token-level word translation. We evaluate the translation performance of visual entities (such as "cars," "clothes" and "bicycles") and poses (e.g., "hold") on a randomly selected set of 100 samples. The number of semantically correct translated words are listed in Table X.

Table X shows that 1) the number of correct translated visual-related words decreases when the original images are replaced with randomly selected images, and 2) incorporating visual information improves the quality of token-level word translation when comparing our model to the text-only NMT

Source:	a man with dark colored hair and a short beard wearing a rust colored tshirt is eating spaghetti at a table .
Reference:	ein mann mit dunklen haaren und einem kurzen bart , der ein rostfarbenes t-shirt trägt und an einem tisch spagetti isst .
NMT:	ein mann mit dunklen haaren und einem kurzen bart in einem hemd , der einen ghetroten oberteil trägt , isst an einem tisch . (**En:** a man with dark hair and a short beard in a shirt wearing a ghee red shirt eats at a table .)
Doubly-ATT:	ein mann mit dunklen haaren und einem kurzen bart in einem rustfarbenen hemd isst spafe an einem tisch . (**En:** a man with dark hair and a short beard in a rust colored shirt eats spafe at a table .)
Multimodal self-att:	ein mann mit dunklen haaren und einem kurzen bart in einem t-shirt isst spaghet-oberteil auf einem tisch . (**En:** a man with dark hair and a short beard in a t-shirt eats spaghet top on a table .)
Gated Fusion:	ein mann mit dunklen haaren und einem kurzen bart , der einen verzierten bart trägt , isst in einem spaghettisch . (**En:** a man with dark hair and a short beard wearing a decorated beard eats at a spaghetti table .)
Our model:	ein mann mit dunklen haaren und einem kurzem bart in einem roten oberteil isst spaghetti an einem tisch . (**En:** a man with dark hair and a short beard in a red top is eating spaghetti at a table .)

Source:	a log cabin is surrounded by trees and bushes .
Reference:	eine holzhütte ist von bäumen und büschen umgeben .
NMT:	eine hütte hütte ist von bäumen und büschen umgeben . (**En**: a hut hut is surrounded by trees and bushes .)
Doubly-ATT:	eine hütte ist von bäumen und büschen umgeben . (**En**: a hut is surrounded by trees and bushes .)
Multimodal self-att:	ein kabinenbaumstamm ist von bäumen und büschen umgeben . (**En**: a cabin log is surrounded by trees and bushes .)
Gated Fusion:	eine hütte ist von bäumen umgeben und büschen . (**En**: a cottage is surrounded by trees and bushes.)
Our model:	eine holzhütte ist von bäumen und büschen umgeben . (**En**: a log cabin is surrounded by trees and bushes .)

Fig. 7. Examples of translations of different MNMT models. Visualize En → De translation results on the Test2017 test set. The improved translation is highlighted in blue.

TABLE IX
AVERAGE FIDELITY SCORE ON THE EN-DE AND EN-FR TRANSLATION TASKS

Multi30K **En-De** Translation Task			
Model	Test2016	Test2017	MSCOCO
Our model (original image-text pair)	83.3	76.5	80.9
Our model (mismatched image-text pair)	75.7	69.1	67.1
Text-only NMT approach	77.4	70.2	72.4
Multi30K **En-Fr** Translation Task			
Model	Test2016	Test2017	MSCOCO
Our model (original image-text pair)	91.5	86.6	80.9
Our model (mismatched image-text pair)	84.9	80.4	67.1
Text-only NMT approach	86.4	81.9	72.4

Fig. 8. The heat map visualization of image-text pair.

TABLE X
THE QUANTITATIVE HUMAN ANALYSIS ON TOKEN-LEVEL WORD TRANSLATION

Model	Number of entities	Number of poses
Total numbers	277	142
Our model (original image-text pair)	262	129
Our model (mismatched image-text pair)	240	105
Text-only NMT approach	247	121

approach. The human analysis of visual-related word translation further confirms the effectiveness of our proposed approach. The quantitative human analysis on sentence-level fidelity scores and token-level translation further confirms the effectiveness of our proposed approach.

V. CONCLUSION

In this article, we proposed an adaptive layer-level progressive multi-modal fusion approach to address the multi-modal fusion problem raised in MNMT. The MDA module is employed to quantify the modality gap in each transformer layer, and the layer-level progressive fusion strategy is adopted based on MDA to gradually incorporate visual information into texts to enhance machine translation performance. Experimental results on three benchmark translation tasks demonstrate the effectiveness and superiority of our proposed method. Ablation studies show that the proposed method could extract helpful visual information to promote visual-to-textual fusion, and the effectiveness of visual features in machine translation is verified by adversarial evaluation. In addition, case study and human analysis further confirm the effectiness of our proposed approach for machine translation.

REFERENCES

[1] D. Elliott, S. Frank, L. Barrault, F. Bougares, and L. Specia, "Findings of the second shared task on multimodal machine translation and multilingual image description," in *Proc. Conf. Mach. Transl.*, 2017, pp. 215–233.
[2] L. Barrault, F. Bougares, L. Specia, C. Lala, D. Elliott, and S. Frank, "Findings of the third shared task on multimodal machine translation," in *Proc. 3rd Conf. Mach. Transl.: Shared Task Papers*, 2018, pp. 304–323.
[3] J. Ye and J. Guo, "Dual-level interactive multimodal-mixup encoder for multi-modal neural machine translation," *Appl. Intell.*, vol. 52, pp. 14194–14203, 2022.

[4] O. Caglayan, P. Madhyastha, L. Specia, and L. Barrault, "Probing the need for visual context in multimodal machine translation," in *Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics: Hum. Lang. Technol.*, 2019, pp. 4159–4170. [Online]. Available: https://aclanthology.org/N19-1422

[5] Y. Yin et al., "A novel graph-based multi-modal fusion encoder for neural machine translation," in *Proc. 58th Annu. Meeting Assoc. Comput. Linguistics*, 2020, pp. 3025–3035.

[6] J. Li, D. Ataman, and R. Sennrich, "Vision matters when it should: Sanity checking multimodal machine translation models," in *Proc. Conf. Empir. Methods Natural Lang. Process.*, 2021, pp. 8556–8562.

[7] S.-A. Grönroos et al., "The MeMAD submission to the WMT18 multimodal translation task," in *Proc. 3rd Conf. Mach. Transl.: Shared Task Papers*, 2018, pp. 603–611. [Online]. Available: https://aclanthology.org/W18-6439

[8] S. Yao and X. Wan, "Multimodal transformer for multimodal machine translation," in *Proc. 58th Annu. Meeting Assoc. Comput. Linguistics*, 2020, pp. 4346–4350.

[9] H. Takushima, A. Tamura, T. Ninomiya, and H. Nakayama, "Multimodal neural machine translation using CNN and transformer encoder," *Adv. Natural Lang. Process.*, p. 85, 2019.

[10] B. Gain, D. Bandyopadhyay, and A. Ekbal, "Experiences of adapting multimodal machine translation techniques for hindi," in *Proc. 1st Workshop Multimodal Mach. Transl. Low Resource Lang.*, 2021, pp. 40–44.

[11] H. Lin et al., "Dynamic context-guided capsule network for multimodal machine translation," in *Proc. 28th ACM Int. Conf. Multimedia*, 2020, pp. 1320–1329.

[12] S. Kwon, B.-H. Go, and J.-H. Lee, "A text-based visual context modulation neural model for multimodal machine translation," *Pattern Recognit. Lett.*, vol. 136, pp. 212–218, 2020.

[13] D. Wang and D. Xiong, "Efficient object-level visual context modeling for multimodal machine translation: Masking irrelevant objects helps grounding," in *Proc. 35th AAAI Conf. Artif. Intell.*, 2021, pp. 2–9.

[14] Y. Song, S. Chen, Q. Jin, W. Luo, J. Xie, and F. Huang, "Enhancing neural machine translation with dual-side multimodal awareness," *IEEE Trans. Multimedia*, vol. 24, pp. 3013–3024, 2022.

[15] Y. Zhao, M. Komachi, T. Kajiwara, and C. Chu, "Word-region alignment-guided multimodal neural machine translation," in *IEEE/ACM Trans. Audio Speech Lang. Process.*, vol. 30, pp. 244–259, 2021.

[16] Y. Zhao, M. Komachi, T. Kajiwara, and C. Chu, "Region-attentive multimodal neural machine translation," *Neurocomputing*, vol. 476, pp. 1–13, 2022.

[17] M. Zhou, R. Cheng, Y. J. Lee, and Z. Yu, "A visual attention grounding neural model for multimodal machine translation," in *Proc. Conf. Empir. Methods Natural Lang. Process.*, 2018, pp. 3643–3653. [Online]. Available: https://aclanthology.org/D18-1400

[18] P.-Y. Huang, F. Liu, S.-R. Shiang, J. Oh, and C. Dyer, "Attention-based multimodal neural machine translation," in *Proc. 1st Conf. Mach. Transl.*, 2016, pp. 639–645.

[19] I. Calixto, Q. Liu, and N. Campbell, "Doubly-attentive decoder for multimodal neural machine translation," in *Proc. 55th Annu. Meeting Assoc. Comput. Linguistics*, 2017, pp. 1913–1924. [Online]. Available: https://aclanthology.org/P17-1175

[20] J. Su et al., "Multi-modal neural machine translation with deep semantic interactions," *Inf. Sci.*, vol. 554, pp. 47–60, 2021.

[21] D. Elliott, "Adversarial evaluation of multimodal machine translation," in *Proc. Conf. Empir. Methods Natural Lang. Process.*,, 2018, pp. 2974–2978. [Online]. Available: https://aclanthology.org/D18-1329

[22] K. D. Chowdhury and D. Elliott, "Understanding the effect of textual adversaries in multimodal machine translation," in *Proc. Beyond Vis. Language: in Tegrating Real-world knowl.*, 2019, pp. 35–40. [Online]. Available: https://aclanthology.org/D19-6406

[23] J. Ive, A. M. Li, Y. Miao, O. Caglayan, P. Madhyastha, and L. Specia, "Exploiting multimodal reinforcement learning for simultaneous machine translation," in *Proc. 16th Conf. Eur. Chapter Assoc. Comput. Linguistics*, 2021, pp. 3222–3233.

[24] I. Calixto, Q. Liu, and N. Campbell, "Incorporating global visual features into attention-based neural machine translation," in *Proc. Conf. Empir. Methods Natural Lang. Process.*, Copenhagen, Denmark, 2017, pp. 992–1003, doi: 10.18653/v1/d17-1105.

[25] O. Caglayan et al., "LIUM-CVC submissions for WMT17 multimodal translation task," in *Proc. 2nd Conf. Mach. Transl.*, Copenhagen, Denmark, 2017, pp. 432–439, doi: 10.18653/v1/w17-4746..

[26] O. Caglayan, L. Barrault, and F. Bougares, "Multimodal attention for neural machine translation," 2016, *arXiv:1609.03976*.

[27] O. Caglayan et al., "Does multimodality help human and machine for translation and image captioning?," in *Proc. 1st Conf. Mach. Transl.*, 2016, pp. 627–633.

[28] J.-B. Delbrouck and S. Dupont, "An empirical study on the effectiveness of images in multimodal neural machine translation," in *Proc. Conf. Empir. Methods Natural Lang. Process.*, 2017, pp. 910–919. [Online]. Available: https://aclanthology.org/D17-1095

[29] Y. Su, K. Fan, N. Bach, C.-C. J. Kuo, and F. Huang, "Unsupervised multimodal neural machine translation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 10482–10491.

[30] T. Nishihara, A. Tamura, T. Ninomiya, Y. Omote, and H. Nakayama, "Supervised visual attention for multimodal neural machine translation," in *Proc. 28th Int. Conf. Comput. Linguistics*, 2020, pp. 4304–4314.

[31] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 770–778.

[32] Y. Su and C.-C. J. Kuo, "On extended long short-term memory and dependent bidirectional recurrent neural network," *Neurocomputing*, vol. 356, pp. 151–161, 2019, doi: 10.1016/j.neucom.2019.04.044.

[33] H. Zhang, M. Cisse, Y. N. Dauphin, and D. Lopez-Paz, "mixup: Beyond empirical risk minimization," in *Proc. 6th Int. Conf. Learn. Representations*, 2018. [Online]. Available: https://openreview.net/forum?id=r1Ddp1-Rb

[34] S. Yun, D. Han, S. J. Oh, S. Chun, J. Choe, and Y. Yoo, "CutMix: Regularization strategy to train strong classifiers with localizable features," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2019, pp. 6023–6032.

[35] H. Guo, Y. Mao, and R. Zhang, "Mixup as locally linear out-of-manifold regularization," in *Proc. AAAI Conf. Artif. Intell.*, 2019, pp. 3714–3722.

[36] L. Sun, C. Xia, W. Yin, T. Liang, P. S. Yu, and L. He, "Mixup-transformer: Dynamic data augmentation for NLP tasks," in *Proc. 28th Int. Conf. Comput. Linguistics*, 2020, pp. 3436–3440.

[37] Y. Wu, D. Inkpen, and A. El-Roby, "Mixup regularized adversarial networks for multi-domain text classification," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2021, pp. 7733–7737.

[38] H. Guo, "Nonlinear mixup: Out-of-manifold data augmentation for text classification," in *Proc. AAAI Conf. Artif. Intell.*, 2020, pp. 4044–4051.

[39] J. Ye, J. Guo, Y. Xiang, K. Tan, and Z. Yu, "Noise-robust cross-modal interactive learning with text2image mask for multi-modal neural machine translation," in *Proc. 29th Int. Conf. Comput. Linguistics*, 2022, pp. 5098–5108.

[40] A. Vaswani et al., "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst.*, 2017, pp. 5998–6008.

[41] D. Elliott, S. Frank, K. Sima'an, and L. Specia, "Multi30 k: Multilingual english-german image descriptions," in *Proc. 5th Workshop Vis. Lang. Hosted 54th Annu. Meeting Assoc. Comput. Linguistics*, Berlin, Germany, 2016, pp. 70–74, doi: 10.18653/v1/w16-3210.

[42] R. Sennrich, B. Haddow, and A. Birch, "Neural machine translation of rare words with subword units," in *Proc. 54th Annu. Meeting Assoc. Comput. Linguistics*, 2016, pp. 1715–1725. [Online]. Available: https://aclanthology.org/P16-1162

[43] Z. Wu, L. Kong, W. Bi, X. Li, and B. Kao, "Good for misconceived reasons: An empirical revisiting on the need for visual context in multimodal machine translation," in *Proc. 59th Annu. Meeting Assoc. Computat. Linguistics and the 11th Int. Joint Conf. Natural Lang. Process.*, 2021, pp. 6153–6166. [Online]. Available: https://aclanthology.org/2021.acl-long.480

[44] J. Ive, P. Madhyastha, and L. Specia, "Distilling translations with visual awareness," in *Proc. 57th Annu. Meeting Assoc. Comput. Linguistics*, 2019, pp. 6525–6538. [Online]. Available: https://aclanthology.org/P19-1653

[45] J. Zhu et al., "Incorporating bert into neural machine translation," in *Proc. Int. Conf. Learn. Representations*, 2020, pp. 1–12.

[46] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, "BLEU: A method for automatic evaluation of machine translation," in *Proc. 40th Annu. Meeting Assoc. Comput. Linguistics*, 2002, pp. 311–318.

[47] M. Denkowski and A. Lavie, "Meteor universal: Language specific translation evaluation for any target language," in *Proc. 9th Workshop Stat. Mach. Transl.*, 2014, pp. 376–380.

[48] I. Calixto, M. Rios, and W. Aziz, "Latent variable model for multi-modal translation," in *Proc. 57th Annu. Meeting Assoc. Comput. Linguistics*, 2019, pp. 6392–6405. [Online]. Available: https://aclanthology.org/P19-1642

[49] H. S. Arslan, M. Fishel, and G. Anbarjafari, "Doubly attentive transformer machine translation," *CoRR*, 2018.

Junjun Guo received the graduation degree from the China University of Petroleum, Qingdao, China, in 2010, and the Doctoral degree from Xi'an Jiaotong University, Xi'an, China, in 2017. He is currently an Associate Professor with the Kunming University of Science and Technology, Kunming,China. His research interests include pattern recognition, multimodal fusion, and machine translation.

Junjie Ye received the master's degree from the Kunming University of Science and Technology, Kunming, China, in 2023. He is currently working toward the Ph.D. degree with Yunnan University, Kunming, China. His research interests include multimodal fusion, machine translation, fractional order, and time series forecasting.

Yan Xiang received the master's degree from Wuhan University, Wuhan, China, in 2003, and the Doctoral degree from the Kunming University of Science and Technology, Kunming, China, in 2022. She is currently an Associate Professor with the Kunming University of Science and Technology. Her research interests include intelligent information processing and affective computing.

Zhengtao Yu received Ph.D. degree in computer application technology from the Beijing Institute of Technology, Beijing, China, in 2005. He is currently a Professor with the School of Information Engineering and Automation, Kunming University of Science and Technology, Kunming, China. His main research interests include natural language processing, information retrieval, and machine learning.